Basic

Machine Learning

Capstone Project examples

School of Information and Communication Technology

Hanoi University of Science and Technology

1. Prediction of apps' rating

- **Problem**: study to build a system that can make accurate prediction about the average rating for an app, using some descriptions about the app.
- **Input**: some descriptions about the app
- **Output**: average rating from users for a given app
- **Method** to be used: Ridge regression or neural network
- **Dataset**: a set of apps and their descriptions in terms of text, each app has a rating collected from App Store.

2. Prediction of hotels' rating

- **Problem**: study to build a system that can make accurate prediction about the rating for a hotel when it has just been launched, using some descriptions about that hotel. The rating belongs to {1*, 2*, 3*, 4*, 5*}.
- **Input**: some descriptions about the hotel
- **Output**: rating for that hotel
- **Method** to be used: Random Forest
- **Dataset**: a set of hotels and their descriptions. The data will be collected from Agoda.com.

3. Users' preference in music

- **Problem**: analyze the preference/interest of online users about music, over demographic/time/sex, …
- **Input**: set of songs/MV, and a set of users and their interactions with the songs/MV
- **Output**: preference, new conclusion/finding, visualization, …
- **Method** to be used: clustering by K-means, classification with Random forest, …
- **Dataset**: set of songs/MV, and a set of users and their interactions with the songs/MV. The data will be collected from youtube.com.

4. Sentiments about Covid-19

- **Problem**: analyze sentiments of online users about the affects of Covid-19
- **Input**: A tweet from Twitter
- **Output**: sentiment detected, new conclusion/finding, visualization, …
- **Method** to be used: clustering by K-means, classification with SVM, …
- **Dataset**: set of tweets from Twitter.

5. Sentiments about products

- **Problem**: analyze sentiments of online users about some products
- **Input**: A sentence/paragraph/document/post
- **Output**: sentiment detected for each product or each aspect, …
- **Method** to be used: SVM, KNN, …
- **Dataset**: set of sentences/paragraphs/documents/posts, each has one/some labels.

6. Connecting users with products

- **Problem**: analyze preference of online users about some products and then suggest a suitable product to a user
- **Input**: Some interactions (views, buys, clicks) of an user in the past
- **Output**: 5 products that the user may be most interested, …
- **Method** to be used: SVM, KNN, …
- **Dataset**: set of interactions (views, buys, clicks) of online users in the past

7. Neoplastic characterization in colonoscopy

- **Problem**: to segment polyp regions and classify them into two classes: neoplastic or non-neoplastic
- **Input**: medical images
- **Output**: semantic segmentation
- **Method** to be used: encoder-decoder based model, Transformer
- **Dataset**: https://bkai.ai/research/bkai-igh-neopolyp-small-a-dataset-for-fine-grained-polyp-segmentation/

8. Domain Adaptation in Object Detection

- **Problem**: to detect ligaments in liquid sprays and generalize the model to other domains
- **Input**: images of the breakup process of liquid sprays
- **Output**: detected objects
- **Method** to be used: adversarial learning
- **Dataset**: deepspray dataset (internal)

9. Domain Adaptation for classification anatomical sites

- **Problem**: to classify anatomical sites based on images captured in different color modes
- **Input**: medical images
- **Output**: predicted labels
- **Method** to be used: CNN, adversarial learning
- **Dataset**: internal

10. Visual Question Answering

- **Problem**: to answer questions about the contents of images
- **Input**: an image and a question
- **Output**: answer to the question
- **Method** to be used: attention, RNN, transformer
- **Dataset**: https://visualqa.org/

11. Traffic flow forecasting

- **Problem**: to forecast traffic flow
- **Input**: time series of traffic flow
- **Output**: predicted values in the future
- **Method** to be used: Attention, RNN, GCN
- **Dataset**: PeMSD4 and PeMSD8

12. Ventilator Pressure Prediction

- **Problem**: to predict ventilator pressure
- **Input**: time series of ventilator pressure
- **Output**: predicted pressure
- **Method** to be used: Attention, RNN, GCN...
- **Dataset**: https://www.kaggle.com/c/ventilator-pressure-prediction